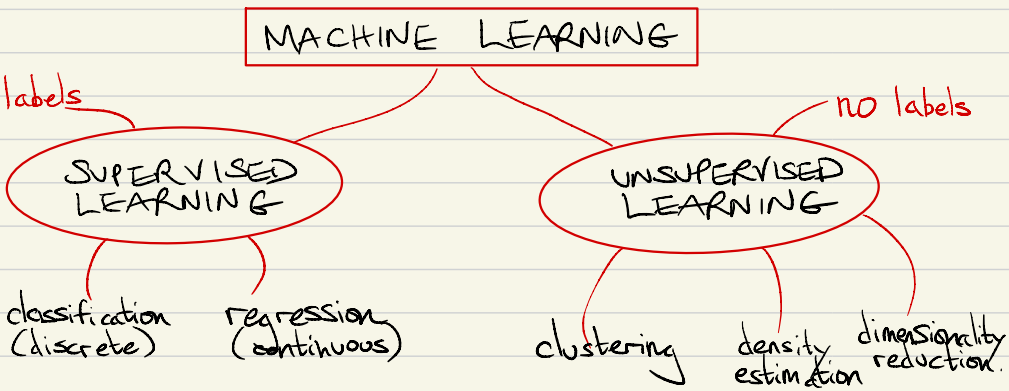
Lecture 13

03/01/2022

Intro to Scikit-Learn



Terminology

* Data recorded in matrix form
  - $[X] = [N_{samples}, N_{features}]$
    - $N_{samples}$: # datapoints
    - $N_{features}$: # attributes

* Some data will have labels (sklearn calls these "TARGETS") — stars, galaxies etc.

- Scikit-learn Workflow

1. Instantiate an estimator object
2. Fit estimator on data and labels
3. Predict new labels
   or
4. Find model parameters

⇒ Usually partition dataset into TRAINING and TESTING sets.

- Supervised learning — labels

* Classification algorithms are trained on labelled data, and used to classify new object features.

* Regression (or "fitting") is the continuous form of classification.

- Unsupervised learning — no labels.

* Use the data to discover its own labels.

* Clustering (group similar data), density estimation (find the PDF), dimensionality reduction (find important features).